Accelerated Nonparametric Maximum Likelihood Density Deconvolution Using Bernstein Polynomial

Zhong Guan

Department of Mathematical Sciences

Indiana University South Bend, South Bend, IN 46634

December 6, 2016

Abstract

A new maximum likelihood method for deconvoluting a continuous density with a positive lower bound on a known compact support in additive measurement error models with known error distribution using the approximate Bernstein type polynomial model, a finite mixture of specific beta distributions, is proposed. The change-point detection method is used to choose an optimal model degree. Based on a contaminated sample of size n, under an assumption which is satisfied, among others, by the generalized normal error distribution, the optimal rate of convergence of the mean integrated squared error is proved to be $k^{-1}\mathcal{O}(n^{-1+1/k}\log^3 n)$ if the underlying unknown density has continuous $2k$th derivative with $k > 1$. Simulation shows that small sample performance of our estimator is better than the deconvolution kernel density estimator. The proposed method is illustrated by a real data application.

Key Words and Phrases: Bernstein Polynomial Model, Beta mixture model; Deconvolution; Density estimation; Kernel density; Measurement error model; Model selection.

1 Introduction

Let X and ε be independent random variables. We are interested in estimating the probability density function f of X. However, in real world, due to measurement error ε with known density g, we have only n independent and identically distributed observations $y_1, \ldots, y_n$ having the same density ψ as that of $Y = X + \varepsilon$ in the additive measurement error model. The data contaminated with additive measurement errors are common in application of statistics. A simple example is the round-off errors with known uninform distribution on $[-0.5/10^d, 0.5/10^d]$ for some integer d. Usually in this case the errors are ignored if d is large. However, in some situations, ignoring the measurement errors can result in serious bias in statistical inference.

In the additive measurement error model, ψ is the convolution of f and g, i.e., $\psi(y) = (f * g)(y) = \int g(y-x)f(x)dx$. So the contaminated data $y_1, \ldots, y_n$ is a sample from a compound distribution ψ or a mixture of the translated $g(y-x)$ with unknown mixing density $f(x)$. Based on the contaminated data a nonparametric estimator $\hat{f}_{\mathrm{F}}$, also known as deconvolution kernel density estimator, of f (see Carroll and Hall, 1988; Devroye, 1989; Stefanski and Carroll, 1990, for examples) is obtained by the inverse Fourier transform with the aid of the kernel density estimation. Briefly, let $\hat{\psi}_{\mathrm{K}}$ be a kernel density estimate of ψ based on $y_1, \ldots, y_n$. Let $\mathcal{F}(\varphi)$ denote the Fourier transform of φ. Since $\mathcal{F}(\psi) = \mathcal{F}(g)\mathcal{F}(f)$, one can estimate $\mathcal{F}(f)$ by $\mathcal{F}(\hat{\psi}_{\mathrm{K}})/\mathcal{F}(g)$ and obtain an $\hat{f}_{\mathrm{F}}$ by inverse Fourier transform.

The properties of the above deconvolution method have been extensively studied by, among others, Zhang (1990); Fan (1991, 1992); Efromovich (1997). Delaigle and Meister (2008) considered the kernel density deconvolution with heteroscedastic errors, i.e., ε_i's have different densities g_i. The optimal rates of convergence for nonparametric deconvolution are extremely slow especially for supersmooth error distributions including normal distribution (see Carroll and Hall, 1988; Fan, 1991, 1992, for example). Specifically, if errors have a super-smooth error distribution such as normal distribution and f satisfies some smooth conditions but without assuming a compact support and a positive lower bound, then the optimal convergence rate of the pointwise mean squared error for a nonparametric estimator of f based on the contaminated data $y_1, \ldots, y_n$ is $\mathcal{O}\{(\log n)^{-\eta}\}$ for some $\eta > 0$ which can be attained by a kernel density estimator $\hat{f}_{\mathrm{F}}$. Although it has been shown by Fan (1992) that nonparametric deconvolution

with normal errors can be as good as the kernel density estimate based on uncontaminated data if the noise level is not too high, an accelerated denconvolution is still much desirable. Recently Delaigle and Hall (2014) proposed an improved kernel method upon $\hat{f}_{\mathrm{F}}$ to speed up the convergence assisted by a "close to being correct" parametric guess of f .

Without measurement errors the kernel density $\tilde{f}_{\mathrm{K}}$ has expectation $\mathrm{E}\{\tilde{f}_{\mathrm{K}}(x)\} = \int K_h(x-y)f(y)dy = (K_h * f)(x)$. So $\tilde{f}_{\mathrm{K}}$ is an "unbiased" estimator of the convolution of f and the scaled kernel $K_h(\cdot) = K(\cdot/h)/h$. No matter how the kernel K and the bandwidth h are chosen, there is always trade-off between the bias and the variance. In this context, Guan (2015) proposed a new nonparametric maximum likelihod estimate for a density which is assumed to be smooth function with a positive lower bound on a known compact support. This method approximately parametrizes the underlying density f, after truncation and transformation to $[0,1]$, by the Bernstein type polynomials which is actual a mixture of beta densities β_{mi} $(i = 0,\ldots,m)$ with shape parameters $(i+1, m-i+1)$, i.e., $f_m(x, \boldsymbol{p}_m) = \sum_{i=0}^{m} p_i \beta_{mi}(x)$. Guan (2015) suggested a change-point detection method to choose an optimal degree m. It has been shown that this new estimate enjoys an almost parametric rate of convergence in mean integrated squared error. Therefore under the same assumptions an accelerated density deconvolution by using the Bernstein polynomial density estimation can be expected. The assumption of a known error density g was discussed by Horowitz and Markatou (1996); Efromovich (1997); Neumann (1997); Efromovich (1999).

We will show that in the additive measurement error model the convolution density ψ can be approximated by a mixture model of known components but unknown mixture proportions. Consequently, we can deconvolute for f using an approximate maximum likelihood method. The resulting density estimate could attain a much better convergence rate. This method is totally different from those in the literature. It does not use the Fourier transforms and can be viewed as a nearly parametric approach to the nonparametric density deconvolution. Like any finite mixture model, this approximate model is different from the classical parametric models because the number of the parameters, the degree of the polynomial, is unknown.

2 Main Results

2.1 Mathematics Preparation

Assume that the density f is continuous on its support $[0,1]$. Then we have (Bernstein, 1912, 1932) $f(u) \approx B_m(f)(u) = \sum_{i=0}^m p_i\beta_{mi}(u)$, where $p_i = f(i/m)/(m+1)$ $(i = 0,\ldots,m)$. The best degree of approximation of f by $B_m(f)$ is $\mathcal{O}(m^{-1})$ no matter how smooth f is. Let $C^{(r)}[0,1]$ be the class of functions which have rth continuous derivative $f^{(r)}$ on $[0,1]$. We denote the m-simplex by $\mathbb{S}_m = \big\{\boldsymbol{p}_m = (p_0,\ldots,p_m)^{\mathrm{T}} : p_i \geq 0,\ \ \sum_{i=0}^m p_i = 1\big\}$. The Bernstein polynomial model is supported by the following mathematical result which is a consequence of Theorem 1 of Lorentz (1963).

Proposition 1 (Lorentz (1963))**.** *If density $f \in C^{(r)}[0,1]$ and $f(u) \geq c > 0$ on its support $[0,1]$, then there exists a sequence of Bernstein type polynomials $f_m(u;\boldsymbol{p}_m) = \sum_{i=0}^m p_i\beta_{mi}(u)$ with $\boldsymbol{p}_m \in \mathbb{S}_m$, such that*

$$|f(u) - f_m(u;\boldsymbol{p}_m)| \leq C(r,c,f)m^{-r/2} \quad (0 \leq u \leq 1), \tag{1}$$

where $C(r,c,f)$ depends on r, c, $\max_u|f(u)|$, and $\max_u|f^{(i)}(u)|$ $(i = 2,\ldots,r)$ only.

The best approximation is unique(Passow, 1977). So we have a parametric approximate model for an arbitrary density $f \in C^{(r)}[0,1]$ $(r \geq 1)$ with positive lower bound on support $[0,1]$. Thus the density ψ can be approximated by $\psi_m(y;\boldsymbol{p}_m) = (g * f_m)(y) = \sum_{i=0}^m p_i(g*\beta_{mi})(y)$, where $(g*\beta_{mi})(y) = \int_0^1 g(y-x)\beta_{mi}(x)dx$ $(i = 0,\ldots,m)$. Therefore the convolution ψ is approximately parameterized as a mixture of $(g*\beta_{mi})$ $(i = 0,\ldots,m)$.

2.2 Maximum Likelihood Estimate

For a given m, the Bernstein likelihood of $y_1,\ldots,y_n$ is defined as $\mathscr{L}(\boldsymbol{p}_m) = \prod_{j=1}^n \sum_{i=0}^m p_i(g*\beta_{mi})(y_j) \approx \prod_{j=1}^n \psi(y_j)$. So the Bernstein loglikelihood is $\ell(\boldsymbol{p}_m) = \sum_{j=1}^n \log\sum_{i=0}^m p_i(g*\beta_{mi})(y_j)$. The maximizer $\hat{\boldsymbol{p}}_m$ of $\ell(\boldsymbol{p}_m)$ is called the maximum Bernstein likelihood estimator of $\boldsymbol{p}_m = (p_0,\ldots,p_m)^{\mathrm{T}}$, the unknown mixture proportions. Then we obtain an estimate of f, $\hat{f}_{\mathrm{B}}(x) = f_m(x;\hat{\boldsymbol{p}}_m)$, for an optimal degree m. The consequent density estimator $\hat{f}_{\mathrm{B}}(x)$ is an approximately parametric density estimator. So it is not

surprising that $\hat{f}_{\mathrm{B}}$ performs much better than a totally nonparametric density estimator such as kernel density estimators which do not take advantage of the conditions imposed on f in this paper.

The expectation-maximization algorithm (Dempster et al., 1977; Wu, 1983; Redner and Walker, 1984) applies to find $\hat{\boldsymbol{p}}_m$ and leads to the following simple iteration:

$$p_l^{(s+1)} = \frac{1}{n}\sum_{j=1}^{n} \frac{p_l^{(s)}(g * \beta_{ml})(y_j)}{\sum_{i=0}^{m} p_i^{(s)}(g * \beta_{mi})(y_j)} \quad (l = 0, \ldots, m;\ s = 0, 1, \ldots). \tag{2}$$

Redner and Walker (1984) proved the convergence of $\boldsymbol{p}_m^{(s)} = (p_0^{(s)}, \ldots, p_m^{(s)})^{\mathrm{T}}$ to $\hat{\boldsymbol{p}}_m$ as $s \to \infty$.

If f is continuous on a support S different from [0, 1] and we can find a finite interval $[a, b] \subset S$ such that $[y_{(1)}, y_{(n)}] \subset [a, b]$ and $F(b) - F(a) \approx 1$, then we let $y_j^* = (y_j - a)/(b - a) = x_j^* + \varepsilon_j^*$, where $x_j^* = (x_j - a)/(b - a)$ and $\varepsilon_j^* = \varepsilon_j/(b - a)$. The densities of x_j^* and ε_j^* are $f^*(x) = (b-a)f\{a + (b-a)x\}$ and $g^*(x) = (b-a)g\{(b-a)x\}$ respectively. Let $\hat{f}_{\mathrm{B}}^*$ be an estimate of f^* based on y_j^*'s. Then we can estimate f by $\hat{f}_{\mathrm{B}}(x) = \hat{f}_{\mathrm{B}}^*\{(x - a)/(b - a)\}/(b - a)$. Since the error distribution is known, we can choose (a, b) by properly extending $(y_{(1)}, y_{(n)})$. Because $x_j = y_j - \varepsilon_j$, we can choose $(a, b) = (y_{(1)} - \zeta\sigma_\varepsilon, y_{(n)} + \zeta\sigma_\varepsilon)$, for some $\zeta > 0$, where σ_ε is the standard deviation of the error ε.

2.3 Model Degree Selection

Denote the sample mean and variance of $y_1, \ldots, y_n$, respectively, by $\bar{y}$ and s^2. Since $\mu_0 = \mathrm{E}(\varepsilon) = 0$ and $\sigma_0^2 = \mathrm{E}(\varepsilon^2)$ are known, we can estimate $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$ by $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = s^2 - \sigma_0^2$, respectively. As in Guan (2015) we can estimate the lower bound m_b for m by $\hat{m}_b = \max\{\lceil \hat{\mu}(1 - \hat{\mu})/\hat{\sigma}^2 - 3 \rceil, 1\}$. Based on $\hat{m}_b$ we choose an appropriate m_0 and a large positive integer I to form $\mathcal{M} = \{m_i = m_0 + i,\ i = 0, \ldots, I\}$. Denote $\ell_i = \ell(\hat{\boldsymbol{p}}_{m_i})$ $(i = 1, \ldots, I)$. Because $f_m(u; \boldsymbol{p}_m)$ is nested in $f_{m+1}(u; \boldsymbol{p}_{m+1})$ (Guan, 2015), so is $\psi_m(u; \boldsymbol{p}_m)$ in $\psi_{m+1}(u; \boldsymbol{p}_{m+1})$. Thus $u_i = \ell_i - \ell_{i-1}$ $(i = 1, \ldots, I)$ are nonnegative. From some real data analysis and extensive simulation study we learned that for large I the optimal degree m_q corresponds such a change-point q that $\{u_{q+1}, \ldots, u_I\}$ have smaller mean and variance than $\{u_1, \ldots, u_q\}$. We can treat $u_1, \ldots, u_I$ as they were exponential observations. The change-point q can be estimated (see §1.4 of Csörgő and Horváth, 1997) by $\hat{q} = \arg\max_{1 \le q < I}\{R(q)\}$, where $R(q) = I \log\{(\ell_I - \ell_0)/I\} - q \log\{(\ell_q - \ell_0)/q\} - (I - q)\log\{(\ell_I - \ell_q)/(I - q)\}$. Having obtained

$\hat{\boldsymbol{p}}_m = (\hat{p}_0, \ldots, \hat{p}_m)^{\mathrm{T}}$, we can use $p_i^{(0)} = \{i\hat{p}_{i-1} + (m-i+1)\hat{p}_i\}/(m+1)$ $(i = 0, \ldots, m+1)$ as the initial guess for the iteration (2) for $\hat{\boldsymbol{p}}_{m+1}$.

3 Asymptotic Results

We will show our asymptotic results assuming $f_m(u; \boldsymbol{p}) = \sum_{i=0}^m p_i \beta_{mi}(u)$ as an approximate model instead of an exact parametric model. For a given $\boldsymbol{p}_0$, we define

$$\|\boldsymbol{p} - \boldsymbol{p}_0\|_g^2 = \int_{-\infty}^{\infty} \frac{\{\psi_m(y; \boldsymbol{p}) - \psi_m(y; \boldsymbol{p}_0)\}^2}{\psi_m(y; \boldsymbol{p}_0)} I\{\psi_m(y; \boldsymbol{p}_0) > 0\} dy.$$

It is clear that $\|\boldsymbol{p} - \boldsymbol{p}_0\|_g^2 = (\boldsymbol{p} - \boldsymbol{p}_0)^{\mathrm{T}} \boldsymbol{C}_m(g)(\boldsymbol{p} - \boldsymbol{p}_0)$, where the square matrix $\boldsymbol{C}_m(g) = \boldsymbol{C}_m(g, \boldsymbol{p}_0) = [\boldsymbol{C}_m^{(ij)}(g, \boldsymbol{p}_0)]_{\{0 \le i,j \le m\}}$ has entries

$$\boldsymbol{C}_m^{(ij)}(g, \boldsymbol{p}_0) \quad = \quad \int_{-\infty}^{\infty} \frac{\int_0^1 \int_0^1 g(y-u) g(y-v) \beta_{mi}(u) \beta_{mj}(v) du dv}{\int_0^1 g(y-x) f_m(x; \boldsymbol{p}_0) dx} dy \quad (i, j = 0, \ldots, m).$$

We need the following assumptions for the asymptotic properties of $\hat{f}_{\mathrm{B}}$ which will be proved in the appendix:

Assumption 1. *$f \in C^{(2k)}[0,1]$ for some $k \ge 1$, and $f(x) \ge c > 0$ on its support $[0,1]$.*

Assumption 2. *$\int_0^1 [\int_{-\infty}^{\infty} g(z) \{\log \psi_m(x+z; \boldsymbol{p}_m)\}^2 dz] dx \le C$ for all $\boldsymbol{p}_m \in \mathbb{S}_m$ and $m > 0$.*

The generalized normal distribution has density $g(x) = g(x; \alpha, \gamma) = \gamma\{2\alpha\Gamma(1/\gamma)\}^{-1} e^{-(|x|/\alpha)^\gamma}$ $(-\infty < x < \infty)$, where $\alpha, \gamma > 0$, and $\Gamma(s)$ is the gamma function. We have the following result.

Proposition 2. *The generalized normal density $g(\cdot) = g(\cdot; \alpha, \gamma)$ satisfies Assumption 2.*

The generalized normal distribution has mean zero and variance $\sigma^2 = \alpha^2 \Gamma(3/\gamma)/\Gamma(1/\gamma)$. Special cases are the supper smooth normal distribution $\mathrm{N}(0, \sigma^2)$ with $(\alpha, \gamma) = (\sqrt{2}\sigma, 2)$ and the ordinary smooth Laplace distribution $\mathrm{L}(0, \sigma)$ with mean 0, variance σ^2, and $(\alpha, \gamma) = (\sigma/\sqrt{2}, 1)$. As $\gamma \to \infty$, $g(x; \alpha, \gamma)$ converges to the uniform $(-\alpha, \alpha)$.

Theorem 1. *Under Assumptions 1 and 2, as $n \to \infty$, with probability one the maximum value of $\ell(\boldsymbol{p}_m)$ with $m = \mathcal{O}(n^{1/k})$ is attained at $\hat{\boldsymbol{p}}_m$ in the interior of $\mathbb{B}_m(r_n) = \{\boldsymbol{p} \in \mathbb{S}_m : \|\boldsymbol{p} - \boldsymbol{p}_m^{(0)}\|_g^2 \le r_n^2\}$,*

where $r_n = n^{-1/2}\log n$ and $\boldsymbol{p}_m^{(0)}$ makes $f_m(\cdot;\boldsymbol{p}_m^{(0)})$ the unique best approximation and the mean weighted integrated squared error of $\hat\psi(y) = \psi_m(y;\hat{\boldsymbol{p}}_m)$ satisfies

$$\mathrm{E}\int \frac{\{\psi_m(y;\hat{\boldsymbol{p}}_m) - \psi(y)\}^2}{\psi(y)}dy = \mathcal{O}\left(\frac{\log^2 n}{n}\right). \tag{3}$$

Remark 1. *If $g = \delta$, the Dirac delta, which satisfies Assumption 2, then under Assumption 1, with $m = \mathcal{O}(n^{1/k})$, (3) is true for $\psi = f$ and $\psi_m = f_m$.*

As a consequence of Theorem 1 and a necessary condition for maximum likelihood estimator $\hat{\boldsymbol{p}}$ (see (3·8) on Page 209 of Redner and Walker, 1984) we have the following much faster convergence rate of $\hat f_{\mathrm{B}}$ than that of $\hat f_{\mathrm{F}}$.

Theorem 2. *Under the conditions of Theorem 1 with $k > 1$ the mean integrated squared error of $\hat f_{\mathrm{B}}$ satisfies*

$$\mathrm{E}\int_0^1 \{f_m(y;\hat{\boldsymbol{p}}_m) - f(y)\}^2 dy = \frac{1}{k}\mathcal{O}(n^{1/k-1}\log^3 n). \tag{4}$$

4 Simulation

In order to exam the finite sample performance of the proposed method, we conduct simution studies to compare the estimator $\hat f_{\mathrm{B}}$ with the surreal parametric deconvolution $\hat f_{\mathrm{P}}$, the Fourier transform estimator $\hat f_{\mathrm{F}}$, and the kernel density $\tilde f_{\mathrm{K}}$ based on the uncontaminated simulated data $x_1,\ldots,x_n$. As in Fan (1992) we generated samples $x_1,\ldots,x_n$ of size $n = 100, 200, 400$ from two distributions: (i) unimodal $\mathrm{N}(0,1)$ truncated by $[a,b] = [-7,7]$, and (ii) bimodal $0{\cdot}6\mathrm{N}(-2,1^2) + 0.4\mathrm{N}(2,0{\cdot}8^2)$ truncated by $[a,b] = [-7,7]$. The errors ε were generated from normal $\mathrm{N}(0,\sigma_0^2)$ and $\mathrm{L}(0,\sigma_0)$, with $\sigma_0 = 0$2$, 0$4$, 0$6$, 0$8$, 1$0. Only when σ_0, compared with the standard deviation σ of X, is "small" one can obtain an applicable estimate of f even for parametric deconvolution. For instance, if both X and ε are normal, then the maximum likelihood estimates of $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$ are, respectively, $\hat\mu = \bar y$ and $\hat\sigma^2 = \max\{(n-1)s^2/n - \sigma_0^2, 0\}$, where $\bar y$ and s^2 are the sample mean and sample variance of $y_1,\ldots,y_n$. If $\sigma < \sigma_0$ and n is not large, then $\hat\sigma^2$ could also be zero because $\Pr\{(n-1)s^2/n < \sigma_0^2\} = \Pr[\chi^2_{n-1} < n/\{1+(\sigma/\sigma_0)^2\}]$ is not small. So the parametric deconvolution $\hat f_{\mathrm{P}}(x)$ may be degenerate because $\hat\sigma^2$ could be zero even if $\sigma_0 \le \sigma$. In the simulation shown in Table 1, $\hat f_{\mathrm{P}}(x)$ is the parametric estimate of the density of $\mathrm{N}(\mu,\sigma^2)$ and λ

$N(\mu_1, \sigma_1^2) + (1-\lambda)N(\mu_2, \sigma_2^2)$ with known variances σ^2, σ_1^2 and σ_2^2 but unknown μ and (λ, μ_1, μ_2). The rate of convergence in mean integrated squared error of such parametric estimator is $\mathcal{O}(n^{-1})$.

The (mixture) normal density f has continuous k-th derivative $f^{(k)}$ for all k. In real world problem, a random variable may just have an approximate normal distribution supported by the central limit theorem and some goodness-of-fit test. In another simulation study presented by Table 2, we generated sample $x_1, \ldots, x_n$ from a "nearly normal" distribution NN(d), $d = 4$, which is the distribution of the sample mean of $u_1, \ldots, u_d$ from uniform(0, 1). Clearly $\mathrm{NN}(d) \approx \mathrm{N}(1/2, 1/(12d))$ for large d. Let f_d be the density of NN(d). If $d > 1$, then $f_d \in C^{(d-2)}[0,1]$ but $f_d \notin C^{(d-1)}[0,1]$. Let $\rho(n)$ denote the probability that the Shapiro-test based on a sample of size n rejects the normality of f_4 with significance level 0·05. Based on 5,000 Monte Carlo runs $\rho(n)$ is estimated to be 0·0398, 0·0504, and 0·0966, respectively, for $n = 100, 200$, and 400. The parametric estimate $\hat{f}_{\mathrm{P}}$ in this simulation is based on the normal model $\mathrm{N}(\mu, \sigma^2)$ with known $\sigma^2 = 1/(12d)$. The errors $\varepsilon_1, \ldots, \varepsilon_n$ were generated from $\mathrm{N}(0, \sigma_0^2)$ and $\mathrm{L}(0, \sigma_0)$, where $12d\sigma_0^2 = 0{\cdot}2^2, 0{\cdot}4^2, 0{\cdot}6^2, 0{\cdot}8^2, 1{\cdot}0^2$ and $d = 4$. So $\sigma_0 = (0{\cdot}05, 0{\cdot}10, 0{\cdot}15, 0{\cdot}20, 0{\cdot}25)/\sqrt{3}$. Although $f_4 \in C^{(2k)}[0,1]$ $(k \leq 1)$ and the condition $k > 1$ of Theorem 2 is not fulfilled, the proposed estimator $\hat{f}_{\mathrm{B}}$ still performs better than $\hat{f}_{\mathrm{F}}$ in such bad scenario. In this case $\hat{f}_{\mathrm{B}}$ performs worse than $\tilde{f}_{\mathrm{K}}$ because the latter is based on uncontaminated data and has a rate of $\mathcal{O}(n^{-4/5})$.

We used the R package decon (Wang and Wang, 2011) which implements the methods of Fan (1991, 1992); Delaigle and Gijbels (2004) and Delaigle and Meister (2008) for calculating $\hat{f}_{\mathrm{F}}$. The "dboot2" method was used for choosing an optimal bandwidth h (see Delaigle and Gijbels, 2004; Wang and Wang, 2011, for details).

For an estimator $\hat{f}$, after R Monte Carlo runs, we obtained estimates $\hat{f}^{(1)}, \ldots, \hat{f}^{(R)}$. We then approximate the point-wise mean squared error at $x \in [a,b]$, $\mathrm{pMSE}\{\hat{f}(x)\} = \mathrm{E}\{\hat{f}(x) - f(x)\}^2 = \mathrm{Var}\{\hat{f}(x)\} + \mathrm{Bias}^2\{\hat{f}(x)\}$, by $\widehat{\mathrm{pMSE}}\{\hat{f}(x)\} = \hat{\sigma}^2\{\hat{f}(x)\} + [\hat{\mu}\{\hat{f}(x)\} - f(x)]^2$, where $\hat{\mu}\{\hat{f}(x)\}$ and $\hat{\sigma}^2\{\hat{f}(x)\}$ are the sample mean and the sample variance of $\hat{f}^{(1)}(x), \ldots, \hat{f}^{(R)}(x)$. In order to compare in details the proposed estimator $\hat{f}_{\mathrm{B}}$ with $\hat{f}_{\mathrm{F}}$, $\hat{f}_{\mathrm{P}}$ and $\tilde{f}_{\mathrm{K}}$, we plot the point-wise mean squared error in Figure 1 from which we see that $\hat{f}_{\mathrm{B}}$ almost uniformly outperforms $\hat{f}_{\mathrm{F}}$ for both unimodal and bimodal f. We also see that if f is unimodal and smooth enough so that k is large as in Theorem 2 then $\hat{f}_{\mathrm{B}}$ even almost uniformly outperforms $\tilde{f}_{\mathrm{K}}$ which is based on uncontaminated data. The mean inte-

Table 1: The estimated square root of the MISE multiplied by 100, $100[\widehat{\mathrm{MISE}}(\hat{f})]^{1/2}$, based on 1000 Monte Carlo runs with $x_1, \ldots, x_n$ being generated from normal and mixture normal distributions and errors $\varepsilon_1, \ldots, \varepsilon_n$ from the normal $\mathrm{N}(0, \sigma_0^2)$ and Laplace $\mathrm{L}(0, \sigma_0)$. In the parametric models all the variances are assumed to be known. $M = \{10, 11, \ldots, 100\}$

f	$\mathrm{N}(0,1)$					$0.6\mathrm{N}(-2,1)+0.4\mathrm{N}(2,0.8^2)$				
σ_0	0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
$\varepsilon \sim \mathrm{N}(0,\sigma_0^2)$										
	$n=100$					$n=100$				
$\hat{f}_\mathrm{P}$	3.96	4.08	4.44	4.93	5.35	5.97	6.42	6.83	7.64	8.97
$\hat{f}_\mathrm{B}$	5.26	5.54	5.79	6.24	6.94	7.36	7.82	8.43	9.34	10.93
$\hat{f}_\mathrm{F}$	9.40	11.03	13.13	15.90	19.23	9.10	14.74	16.42	17.98	19.37
$\tilde{f}_\mathrm{K}$	8.14	8.37	8.29	8.21	8.19	8.26	8.35	8.27	8.11	8.26
	$n=200$					$n=200$				
$\hat{f}_\mathrm{P}$	2.72	2.93	3.17	3.48	3.68	4.09	4.31	4.75	5.45	6.35
$\hat{f}_\mathrm{B}$	3.81	4.07	4.36	4.75	5.08	5.71	6.05	6.60	7.45	8.51
$\hat{f}_\mathrm{F}$	7.47	8.93	11.04	13.80	16.63	6.96	11.56	13.44	15.37	17.20
$\tilde{f}_\mathrm{K}$	6.23	6.27	6.31	6.16	6.24	6.23	6.36	6.23	6.26	6.34
	$n=400$					$n=400$				
$\hat{f}_\mathrm{P}$	1.87	2.04	2.12	2.47	2.64	2.96	2.75	3.46	3.81	4.45
$\hat{f}_\mathrm{B}$	2.65	2.90	3.14	3.53	3.83	4.74	4.70	5.46	6.05	6.88
$\hat{f}_\mathrm{F}$	5.90	7.25	9.36	11.91	14.69	5.56	8.84	11.17	13.36	15.42
$\tilde{f}_\mathrm{K}$	4.77	4.72	4.66	4.70	4.71	4.79	4.63	4.84	4.74	4.77
$\varepsilon \sim \mathrm{L}(0,\sigma_0)$										
	$n=100$					$n=100$				
$\hat{f}_\mathrm{P}$	4.24	4.36	5.01	5.68	5.97	6.02	6.55	7.59	8.92	10.49
$\hat{f}_\mathrm{B}$	5.47	5.71	6.58	7.39	8.27	7.41	7.99	9.42	10.61	12.24
$\hat{f}_\mathrm{F}$	14.17	19.17	24.17	27.66	30.59	12.92	16.76	20.42	23.56	26.91
$\tilde{f}_\mathrm{K}$	8.48	8.21	8.36	8.37	7.98	8.15	8.22	8.25	8.18	8.17
	$n=200$					$n=200$				
$\hat{f}_\mathrm{P}$	2.78	3.14	3.62	3.78	4.51	4.26	4.71	5.46	6.38	7.56
$\hat{f}_\mathrm{B}$	3.87	4.30	5.00	5.44	6.38	5.88	6.41	7.28	8.34	9.64
$\hat{f}_\mathrm{F}$	11.69	17.39	22.97	27.59	31.42	10.32	15.03	18.19	22.40	25.94
$\tilde{f}_\mathrm{K}$	6.23	6.09	6.12	6.33	6.23	6.34	6.24	6.34	6.31	6.26
	$n=400$					$n=400$				
$\hat{f}_\mathrm{P}$	1.97	2.07	2.35	2.94	2.98	2.97	3.26	3.80	4.45	5.36
$\hat{f}_\mathrm{B}$	2.73	3.06	3.72	4.21	4.69	4.87	5.26	5.91	6.78	7.77
$\hat{f}_\mathrm{F}$	9.48	15.10	20.28	24.85	27.71	8.47	12.26	15.76	19.92	22.62
$\tilde{f}_\mathrm{K}$	4.79	4.64	4.70	4.75	4.58	4.89	4.81	4.84	4.79	4.82

Table 2: The estimated square root of the MISE multiplied by 100, $100[\widehat{\mathrm{MISE}}(\hat{f})]^{1/2}$, based on 1000 Monte Carlo runs with $x_1, \dots, x_n$ being generated from the nearly normal distribution NN(4) and errors $\varepsilon_1, \dots, \varepsilon_n$ from the normal $\mathrm{N}(0, \sigma_0^2)$ and Laplace $\mathrm{L}(0, \sigma_0)$. We assume the normal distribution $\mathrm{N}(\mu, \sigma^2)$ with known variance $\sigma^2 = 1/48$ as the parametric model. $M = \{2, 3, \dots, 100\}$

	$X \sim \mathrm{NN}(4),\ \varepsilon \sim \mathrm{N}(0, \sigma_0^2)$					$X \sim \mathrm{NN}(4),\ \varepsilon \sim \mathrm{L}(0, \sigma_0)$				
σ_0	$\frac{0.05}{\sqrt{3}}$	$\frac{0.10}{\sqrt{3}}$	$\frac{0.15}{\sqrt{3}}$	$\frac{0.20}{\sqrt{3}}$	$\frac{0.25}{\sqrt{3}}$	$\frac{0.05}{\sqrt{3}}$	$\frac{0.10}{\sqrt{3}}$	$\frac{0.15}{\sqrt{3}}$	$\frac{0.20}{\sqrt{3}}$	$\frac{0.25}{\sqrt{3}}$
	$n = 100$									
$\hat{f}_{\mathrm{P}}$	10.80	11.22	12.07	12.98	14.11	10.85	11.89	13.89	15.26	16.55
$\hat{f}_{\mathrm{B}}$	23.07	26.24	30.81	36.48	42.65	23.42	28.28	35.98	41.29	47.36
$\hat{f}_{\mathrm{F}}$	24.21	54.71	87.24	95.17	95.71	28.31	34.25	39.92	44.97	49.48
$\tilde{f}_{\mathrm{K}}$	21.70	20.79	21.28	21.21	20.92	21.02	21.51	21.00	21.64	21.06
	$n = 200$									
$\hat{f}_{\mathrm{P}}$	8.34	8.61	9.13	9.94	10.60	8.50	9.34	10.13	11.15	11.73
$\hat{f}_{\mathrm{B}}$	16.49	19.71	24.11	29.13	34.43	17.30	22.41	27.72	33.08	37.86
$\hat{f}_{\mathrm{F}}$	20.09	49.60	80.42	95.17	95.80	21.47	26.65	31.80	37.09	41.22
$\tilde{f}_{\mathrm{K}}$	15.90	15.89	15.94	15.92	15.69	16.04	16.28	16.01	15.96	16.28
	$n = 400$									
$\hat{f}_{\mathrm{P}}$	6.83	7.06	7.42	8.01	8.26	6.86	6.95	7.74	8.43	9.18
$\hat{f}_{\mathrm{B}}$	12.22	14.34	18.20	22.45	38.59	12.82	16.40	21.72	25.71	30.33
$\hat{f}_{\mathrm{F}}$	16.91	45.52	75.44	93.35	95.80	16.08	21.04	26.35	30.40	34.48
$\tilde{f}_{\mathrm{K}}$	12.23	12.13	12.16	12.26	12.27	11.98	12.08	12.13	12.01	12.16

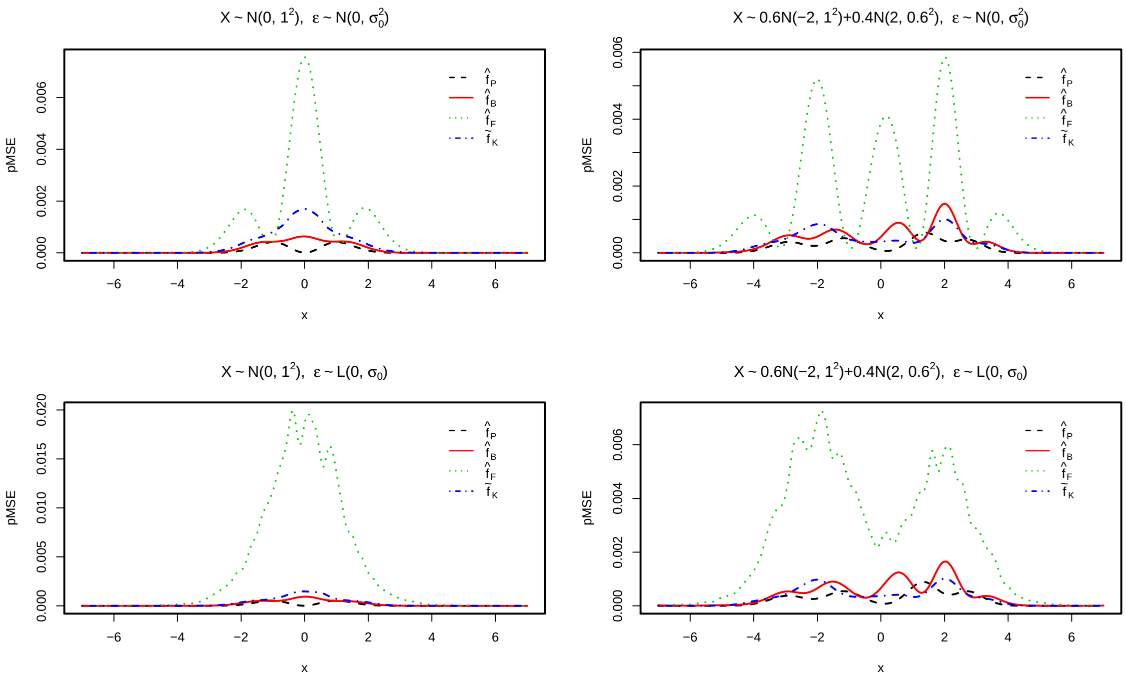

Figure 1: The simulated point-wise MSE's of the parametric estimator $\hat{f}_{\mathrm{P}}$(dashed), the proposed estimator $\hat{f}_{\mathrm{B}}$(solid), the inverse Fourier estimator $\hat{f}_{\mathrm{F}}$(dotted), and the kernel density estimator based on the simulated uncontaminated data $\tilde{f}_{\mathrm{K}}$(dotdash), at $t_i = a + i(b-a)/N$, $i = 0, 1, \ldots, N$, $N = 512$. The sample size is $n = 200$. The truncation interval is $[a, b] = [-7, 7]$. In the parametric models all the variances are assumed to be known. $M = \{10, 11, \ldots, 100\}$. Upper panels: the error distribution is $\mathrm{N}(0, \sigma_0^2)$ with $\sigma_0 = 0.6$; Lower panels: the error distribution is Laplace $\mathrm{L}(0, \sigma_0)$ with $\sigma_0 = 0.6$.

grated squared error $\mathrm{MISE}(\hat{f}) = \mathrm{E}\int[\hat{f}(x) - f(x)]^2dx$ is approximated by $\sum_{i=1}^{N} \mathrm{pMSE}\{\hat{f}(t_i)\}\Delta t$, where $\Delta t = (b-a)/N$, $t_i = a + i\Delta t$ $(i = 0, 1, \ldots, N)$, and $N = 512$. So $\mathrm{MISE}(\hat{f})$ can be estimated by $\widehat{\mathrm{MISE}}(\hat{f}) = \sum_{i=1}^{N} \widehat{\mathrm{pMSE}}\{\hat{f}(t_i)\}\Delta t$.

Tables 1 and 2 show that the proposed $\hat{f}_{\mathrm{B}}$ is much better than the Fourier transform method $\hat{f}_{\mathrm{F}}$. In some cases, especially when σ_0 is much smaller than σ, $\hat{f}_{\mathrm{B}}$ is even as triple efficient as $\tilde{f}_{\mathrm{K}}$ in terms of the square root of the mean integrated squared error. Although the simulation setup prefers the parametric methods the results show that in most cases the proposed approach has mean integrated squared error that leans toward the surreal parametric one. Moreover the proposed method performs better than the kernel estimate based on the uncontaminated data for unimodal model or if the magnitude of error variance is not too large. Because of the involvement of $\tilde{f}_{\mathrm{K}}$ in the comparison it is unnecessary to include any other kernel methods improving upon $\hat{f}_{\mathrm{F}}$ in the simulation.

5 Framingham Data

The Framingham data is from a study on coronary heart disease (Carroll et al., 2006) and consist of measurements of systolic blood pressure in 1,615 males, Y_1 taken at an examination and Y_2 at an 8-year follow-up examination after the first. At the ith examination, the systolic blood pressure was measured twice, Y_{i1} and Y_{i2} $(i = 1, 2)$, for each individual. We used the data in the R package decon (Wang and Wang, 2011) which contain four variables, Y_{ij} $(i, j = 1, 2)$. As in Wang and Wang (2011), the error is assumed to be $N(0, \sigma_0^2)$ with $\sigma_0^2 = 83.69$ estimated using the systolic blood pressures Y_1 at the first examination. The density of the systolic blood pressure Y_2 at the second examination is to be deconvolved. We truncate the distribution by $[a, b] = [80, 270]$ and selected the optimal degree $\hat{m} = 34$ from $\mathcal{M} = \{5, 6, \ldots, 100\}$. Figure 2 shows that the density deconvolutions $\hat{f}_{\mathrm{B}}$, $\hat{f}_{\mathrm{F}}$, and $\tilde{\psi}_{\mathrm{K}}$ ignoring measurement errors are quite different.

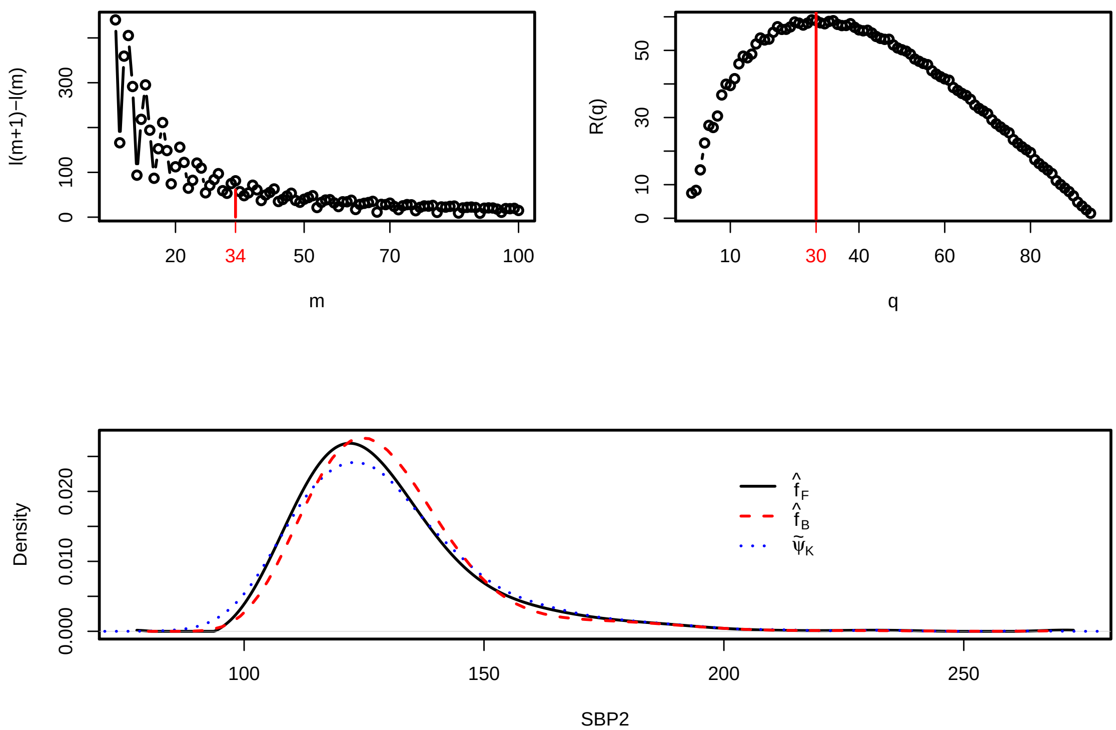

Figure 2: Upper left panel: the increments of loglikelihood vs the model degree $m \in M = \{5, 6, \ldots, 100\}$; Upper right panel: the likelihood ratio of the change point, $\hat{q} = 30$ and $\hat{m} = m_{\hat{q}} = 34$; Lower panel: Density deconvolution of SBP2 based on Framingham data, $\hat{f}_{\mathrm{F}}$ is the inverse Fourier transform estimate(solid); $\hat{f}_{\mathrm{B}}$ is the proposed estimate using Bernstein polynomial with $m = 34$(dashed); $\tilde{\psi}_{\mathrm{K}}$ is the kernel estimate ignoring measurement errors (dotted).

6 Discussion

As shown by Theorem 2 and convinced by the simulation results, the performance of the proposed method leans to that of the parametric approach when the correct parametric model is specified. The classical exact parametric method is subject to model misspecification. Our approach is an approximate parametric solution to a nonparametric problem and speeds up the density deconvolution very much with the computation cost paid for searching an optimal model degree m, of course, under the assumption that the underlying unknown density has a positive lower bound on a known compact support. The condition imposed on the error distribution is satisfied by the family of generalized normal distributions which include the supper smooth normal distribution and the ordinary smooth Laplace distribution.

The technical argument used in the proof of Theorem 1 appears to be new and may be of independent interest. As commented in Remark 1, the special case of this theorem is an enhancement of Theorem 4.1 of Guan (2015) with $g = \delta$ which is proved by the traditional delta method. From the simulation studies we see that there seem rooms for improving the result of Theorem 2.

Acknowledgement

The author would like to thank Professor Jianqing Fan who brought the density deconvolution to his attention.

Appendix

Proof of Proposition 1

Proof. Under the conditions of the proposition, by Theorem 1 of Lorentz (1963) there are polynomials, $\varphi_m(u) = \sum_{i=0}^m c_i \binom{m}{i} u^i (1-u)^{m-i}$ $(m = 0, 1, \ldots)$ with nonnegative coefficients c_i such that $|f(u) - \varphi_m(u)| \le C'(r, c, f)\Delta_m^r(u)$ $(0 \le u \le 1)$, where $\Delta_m(u) = \max[m^{-1}, \{u(1-u)/m\}^{1/2}]$ and $C'(r, c, f)$

depends on r, c, and f only. Since $\Delta_m(u) \le m^{-1/2}$ and $\int f(u)du = 1$, we have $|1 - \sum_{i=0}^m c_i/(m+1)| \le 2C_r M_r m^{-r/2}$. Letting $p_i = c_i / \sum_{i=0}^m c_i$ $(i = 0, \ldots, m)$, then $f_m(u) = \sum_{i=0}^m p_i \beta_{mi}(u)$ satisfies (1). $\square$

Proof of Proposition 2

Proof. It is clear $\log\{\psi_m(y; \boldsymbol{p}_m)\} \le \log\{g(0)\}$. By Jensen's inequality $\log\{\psi_m(y; \boldsymbol{p}_m)\} \ge \sum_{i=0}^m p_i \log\{g * \beta_{mi}(y)\} \ge \sum_{i=0}^m p_i \{(\log \circ g) * \beta_{mi}\}(y) = \sum_{i=0}^m p_i \int_0^1 \log\{g(y-x)\}\beta_{mi}(x)dx$. It is evident

$$\int_0^1 \log\{g(y-x)\}\beta_{mi}(x)dx = \log g(0) - \int_0^1 \left(\frac{|y-x|}{\alpha}\right)^{\gamma} \beta_{mi}(x)dx.$$

Thus by the C_r-inequality

$$\begin{aligned}\{\log[\psi_m(y; \boldsymbol{p}_m)]\}^2 \le & 2|\log g(0)|^2 + 2\sum_{i=0}^m p_i \int_0^1 \left(\frac{|y-x|}{\alpha}\right)^{2\gamma} \beta_{mi}(x)dx \\ \le & 2|\log g(0)|^2 + \frac{2c_{2\gamma}}{\alpha^{2\gamma}}\left(|y|^{2\gamma} + 1\right),\end{aligned}$$

where $c_r = I(0 < r < 1) + 2^{r-1}I(r \ge 1)$. Applying the C_r-inequality again we have

$$\int_0^1 \int_{-\infty}^{\infty} g(z)\{\log \psi_m(x+z; \boldsymbol{p}_m)\}^2 dz dx \le 2|\log g(0)|^2 + \frac{2c_{2\gamma}}{\alpha^{2\gamma}}\left\{c_{2\gamma}\left(\frac{1}{2\gamma+1} + E|\varepsilon|^{2\gamma}\right) + 1\right\}.$$

$\square$

Proof of Theorem 1

Proof. Let $f \in C^{(2k)}[0,1]$ and $f_{m0}(x) = f_m(x; \boldsymbol{p}_m^{(0)})$ be the unique best approximation of degree m for f (Passow, 1977). By Proposition 1 we have $f(x) = f_{m0}(x) + \mathcal{O}(m^{-k})$. Because $f(x) \ge c > 0$, we have, uniformly in m,

$$\rho_m = \sup_{x \in [0,1]} \frac{f_{m0}(x)}{f(x)} = 1 + \mathcal{O}(m^{-k}). \tag{5}$$

Let $u_1, \ldots, u_n$ be iid uniform(0,1) random variables. For each i, if $u_i \le \rho_m^{-1} f_{m0}(x_i)/f(x_i)$, then, by the acceptance-rejection argument used in simulation modeling, x_i $(y_i = x_i + \varepsilon_i)$ can be treated and used as if it were from f_{m0} $(\psi_{m0} = f_{m0} * g)$. Let $A_i = \{u_i \le \rho_m^{-1} f_{m0}(x_i)/f(x_i)\}$ and ν_m be the number

of observations that can be treated and used as if they were from ψ_{m0}. It follows from the law of iterated logarithm that $\nu_m = \sum_{i=1}^n I(A_i) = n - \mathcal{O}(nm^{-k}) - \mathcal{O}\left\{(nm^{-k}\log\log n)^{1/2}\right\}$ almost surely. So we have $\ell(\boldsymbol{p}_m) = \sum_{i=1}^n \log\psi_m(y_i;\boldsymbol{p}_m) = \tilde{\ell}(\boldsymbol{p}_m) + R_{mn}$, where $\tilde{\ell}(\boldsymbol{p}_m) = \sum_{i=1}^n I(A_i)\log\psi_m(y_i;\boldsymbol{p}_m)$ is an "almost complete" likelihood and $R_{mn} = \sum_{i=1}^n I(A_i^c)\log\psi_m(y_i;\boldsymbol{p}_m) \equiv \sum_{i=1}^n W_i$. It is clear that $W_1,\ldots,W_n$ are iid with mean $\mu_W = \mathrm{E}(W_i) = \int_0^1\{f(x)-\rho_m^{-1}f_{m0}(x)\}\int_{-\infty}^{\infty} g(z)\log\{\psi_m(x+z;\boldsymbol{p}_m)\}dzdx$ and variance $\sigma_W^2 = \mathrm{Var}(W_i) = \int_0^1\{f(x)-\rho_m^{-1}f_{m0}(x)\}\int_{-\infty}^{\infty} g(z)\{\log\psi_m(x+z;\boldsymbol{p}_m)\}^2dzdx - \mu_W^2$. By (5) and the conditions of the theorem we have $|\mu_W| \le Cm^{-k}\int_0^1\int_{-\infty}^{\infty} g(z)|\log\psi_m(x+z;\boldsymbol{p}_m)|dzdx = \mathcal{O}(m^{-k})$ and $\sigma_W^2 = \mathcal{O}(m^{-k})$. By the law of iterated logarithm $R_{mn} = \mathcal{O}(nm^{-k}) + \mathcal{O}\left\{(nm^{-k}\log\log n)^{1/2}\right\}$, almost surely. The proportion of the observations that can be treated as if they were from ψ_{m0} is $\nu_m/n = 1 - \mathcal{O}(m^{-k}) - \mathcal{O}\left\{(m^{-k}\log\log n/n)^{1/2}\right\}$, almost surely. Taylor expansions of $\log\psi_m(y_j,\boldsymbol{p})$ at $\log\psi_m(y_j,\boldsymbol{p}_m^{(0)})$ yield that, for $\boldsymbol{p}\in\mathbb{B}_m(r_n)$,

$$\begin{aligned}
\tilde{\ell}(\boldsymbol{p}) &= \sum_{j=1}^n I(A_i)\log\psi_m(y_j,\boldsymbol{p}) \\
&= \tilde{\ell}(\boldsymbol{p}_m^{(0)}) + \sum_{j=1}^n I(A_i)\left[\frac{\psi_m(y_j,\boldsymbol{p})-\psi_m(y_j,\boldsymbol{p}_m^{(0)})}{\psi_m(y_j,\boldsymbol{p}_m^{(0)})} - \frac{1}{2}\frac{\{\psi_m(y_j,\boldsymbol{p})-\psi_m(y_j,\boldsymbol{p}_m^{(0)})\}^2}{\{\psi_m(y_j,\boldsymbol{p}_m^{(0)})\}^2}\right] + \tilde{R}_{mn},
\end{aligned}$$

where $\tilde{R}_{mn} = o(nr_n^2)$, almost surely. Let $\boldsymbol{p}$ be a point on the boundary of $\mathbb{B}_m(r_n)$, i.e., $\|\boldsymbol{p}-\boldsymbol{p}_m^{(0)}\|_g^2 = r_n^2$. By the law of iterated logarithm we have, almost surely,

$$\sum_{j=1}^n I(A_j)\frac{\psi_m(y_j,\boldsymbol{p})-\psi_m(y_j,\boldsymbol{p}_m^{(0)})}{\psi_m(y_j,\boldsymbol{p}_m^{(0)})} = \mathcal{O}\{r_n(n\log\log n)^{1/2}\},$$

and that there exists an $\eta > 0$ such that

$$\sum_{j=1}^n I(A_j)\frac{\{\psi_m(y_j,\boldsymbol{p})-\psi_m(y_j,\boldsymbol{p}_m^{(0)})\}^2}{\{\psi_m(y_j,\boldsymbol{p}_m^{(0)})\}^2} = \eta nr_n^2 + \mathcal{O}\{r_n^2(n\log\log n)^{1/2}\}.$$

Therefore we have, almost surely,

$$\begin{aligned}
\tilde{\ell}(\boldsymbol{p}) &= \tilde{\ell}(\boldsymbol{p}_m^{(0)}) + \sum_{j=1}^n I(A_j)\left[\frac{\psi_m(y_j,\boldsymbol{p})-\psi_m(y_j,\boldsymbol{p}_m^{(0)})}{\psi_m(y_j,\boldsymbol{p}_m^{(0)})} - \frac{1}{2}\frac{\{\psi_m(y_j,\boldsymbol{p})-\psi_m(y_j,\boldsymbol{p}_m^{(0)})\}^2}{\{\psi_m(y_j,\boldsymbol{p}_m^{(0)})\}^2}\right] + o(nr_n^2) \\
&= \tilde{\ell}(\boldsymbol{p}_m^{(0)}) - \frac{1}{2}\eta nr_n^2 + \mathcal{O}\{r_n^2(n\log\log n)^{1/2}\} + \mathcal{O}\{r_n(n\log\log n)^{1/2}\} + o(nr_n^2).
\end{aligned}$$

Since $m = \mathcal{O}(n^{1/k})$, $nm^{-k} = o(nr_n^2)$. So there exists an $\eta' > 0$ such that $\ell(\boldsymbol{p}) \le \ell(\boldsymbol{p}_m^{(0)}) - \eta' nr_n^2 = \ell(\boldsymbol{p}_m^{(0)}) - \eta'(\log n)^2$. Since $\partial^2\ell(\boldsymbol{p})/\partial\boldsymbol{p}\partial\boldsymbol{p}^{\mathrm{T}} < 0$, the maximum value of $\ell(\boldsymbol{p})$ is attained by some $\psi_m(\cdot,\hat{\boldsymbol{p}}_m)$

with $\hat{\boldsymbol{p}}_m$ being in the interior of $\mathbb{B}_m(r_n)$. Then the assertion (3) follows easily from (5). The proof of Theorem 1 is complete. $\square$

Proof of Theorem 2

Proof. Redner and Walker (1984) showed a necessary condition for $\hat{\boldsymbol{p}}$ to maximize $\ell(\boldsymbol{p})$ is

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\hat{p}_i(g*\beta_{mi})(y_j)}{\sum_{l=0}^{m}\hat{p}_l(g*\beta_{ml})(y_j)}-\hat{p}_i=0 \quad (i=0,1,\ldots,m).$$

Therefore we have

$$f_m(x;\hat{\boldsymbol{p}})=\sum_{i=0}^{m}\hat{p}_i\beta_{mi}(x)=\frac{1}{n}\sum_{j=1}^{n}\frac{\sum_{i=0}^{m}\hat{p}_i\beta_{mi}(x)(g*\beta_{mi})(y_j)}{\psi_m(y_j;\hat{\boldsymbol{p}})}.$$

By Taylor expansion we have

$$f_m(x;\hat{\boldsymbol{p}})-f_m(x;\boldsymbol{p}_0)=\frac{1}{n}\sum_{j=1}^{n}\frac{\sum_{i=0}^{m}p_{i0}\beta_{mi}(x)(g*\beta_{mi})(y_j)}{\psi_m(y_j;\boldsymbol{p}_0)}-f_m(x;\boldsymbol{p}_0)+\mathcal{O}\{R_{mn}(x)\},$$

where

$$R_{mn}(x)=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{m}\beta_{mi}(x)(g*\beta_{mi})(y_j)\left[\frac{\hat{p}_i-p_{i0}}{\psi_m(y_j;\boldsymbol{p}_0)}-\frac{p_{i0}\{\psi_m(y_j;\hat{\boldsymbol{p}})-\psi_m(y_j;\boldsymbol{p}_0)\}}{\{\psi_m(y_j;\boldsymbol{p}_0)\}^2}\right].$$

By the law of iterated logarithm and Proposition 1, almost surely,

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\sum_{i=0}^{m}p_{i0}\beta_{mi}(x)(g*\beta_{mi})(y_j)}{\psi_m(y_j;\boldsymbol{p}_0)}-f_m(x;\boldsymbol{p}_0)=\mathcal{O}\{(\log\log n/n)^{1/2}\}.$$

Define, for $0<x<1$,

$$M_m(x) \;=\; \max_{0\le i\le m}\beta_{mi}(x)=(m+1)\binom{m}{\lfloor (m+1)x\rfloor}x^{\lfloor (m+1)x\rfloor}(1-x)^{m-\lfloor (m+1)x\rfloor},$$

where $\lfloor x\rfloor$ is the floor of x. It follows from the Schwartz inequality and Theorem 1 that, almost surely,

$$\begin{aligned}
|R_{mn}(x)|^2 &\le 4M_m^2(x)\Big\{\frac{1}{n}\sum_{j=1}^{n}\frac{|\psi_m(y_j;\hat{\boldsymbol{p}})-\psi_m(y_j;\boldsymbol{p}_0)|}{\psi_m(y_j;\boldsymbol{p}_0)}\Big\}^2\\
&\le 4M_m^2(x)\frac{1}{n}\sum_{j=1}^{n}\frac{\{\psi_m(y_j;\hat{\boldsymbol{p}})-\psi_m(y_j;\boldsymbol{p}_0)\}^2}{\{\psi_m(y_j;\boldsymbol{p}_0)\}^2}\\
&= 4M_m^2(x)\mathcal{O}(n^{-1}\log^2 n).
\end{aligned}$$

By Stirling's approximation we have, for some constants $C > 0$,

$$\begin{aligned}\int_0^1 M_m^2(x)dx &= \sum_{i=0}^{m}\int_{i/(m+1)}^{(i+1)/(m+1)} \beta_{mi}^2(x)dx \\ &\leq (m+1)\left[2+\sum_{i=1}^{m-1}\left\{\binom{m}{i}\left(\frac{i}{m}\right)^i\left(1-\frac{i}{m}\right)^{m-i}\right\}^2\right] \\ &< (m+1)\left(2+C\sum_{i=1}^{m-1}\frac{1}{i}\right) = \frac{1}{k}\mathcal{O}(n^{1/k}\log n).\end{aligned}$$

Thus we have $\mathrm{E}\int\{f_m(x;\hat{\boldsymbol{p}}) - f_m(x;\boldsymbol{p}_0)\}^2dx = \mathcal{O}(n^{-1}\log\log n) + k^{-1}\mathcal{O}(n^{1/k-1}\log^3 n)$. Under the conditions of the theorem, by Proposition 1, we obtain

$$\begin{aligned}\mathrm{E}\int\{f_m(x;\hat{\boldsymbol{p}}) - f(x)\}^2dx &\leq 2\mathrm{E}\int\{f_m(x;\hat{\boldsymbol{p}}) - f_m(x;\boldsymbol{p}_0)\}^2dx + 2\int\{f(x) - f_m(x;\boldsymbol{p}_0)\}^2dx \\ &= \mathcal{O}(n^{-1}\log\log n) + \frac{1}{k}\mathcal{O}(n^{1/k-1}\log^3 n) + \mathcal{O}(n^{-1})\end{aligned}$$

and the proof is complete. □

References

Bernstein, S. N. (1912). "Démonstration du théorème de Weierstrass fondée sur le calcul des probabilitiés." Comm. Soc. Math. Kharkov, vol. 13, pp. 1–2.

— (1932). "Complétement à l'article de E. Voronowskaja." C. R. Acad. Sci. U.R.S.S., pp. 86–92.

Carroll, R. J. and Hall, P. (1988). "Optimal rates of convergence for deconvolving a density." J. Amer. Statist. Assoc., vol. 83, pp. 1184–1186.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). Measurement Error in Nonlinear Models: A Modern Perspective. 2nd ed. New York: Chapman Hall.

Csörgő, M. and Horváth, L. (1997). Limit Theorems in Change-Point Analysis. 1st ed. New York: John Wiley & Sons Inc.

Delaigle, A. and Gijbels, I. (2004). "Bootstrap bandwidth selection in kernel density estimation from a contaminated sample." Ann. Inst. Statist. Math., vol. 56, pp. 19–47.

Delaigle, A. and Hall, P. (2014). "Parametrically assisted nonparametric estimation of a density in the deconvolution problem." Journal of the American Statistical Association, vol. 109, pp. 717–729.

Delaigle, A. and Meister, A. (2008). "Density estimation with heteroscedastic error." Bernoulli, vol. 14, pp. 562–579.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." J. Roy. Statist. Soc. Ser. B, vol. 39, pp. 1–38.

Devroye, L. (1989). "Consistent deconvolution in density estimation." Canad. J. Statist., vol. 17, pp. 235–239.

Efromovich, S. (1997). "Density estimation for the case of supersmooth measurement error." J. Amer. Statist. Assoc., vol. 92, pp. 526–535.

— (1999). Nonparametric Curve Estimation: Methods, Theory, and Applications. Springer Series in Statistics, New York: Springer.

Fan, J. (1991). "On the optimal rates of convergence for nonparametric deconvolution problems." Ann. Statist., vol. 19, pp. 1257–1272.

— (1992). "Deconvolution with supersmooth distributions." Canad. J. Statist., vol. 20, pp. 155–169.

Guan, Z. (2015). "Efficient and robust density estimation using Bernstein type polynomials." Journal of Nonparametric Statistics, to appear.

Horowitz, J. L. and Markatou, M. (1996). "Semiparametric estimation of regression models for panel data." The Review of Economic Studies, vol. 63, pp. pp. 145–168.

Lorentz, G. G. (1963). "The degree of approximation by polynomials with positive coefficients." Math. Ann., vol. 151, pp. 239–251.

Marques, T. A. (2004). "Predicting and correcting bias caused by measurement error in line transect sampling using multiplicative error models." Biometrics, vol. 60, pp. 757–763.

Neumann, M. H. (1997). "On the effect of estimating the error density in nonparametric deconvolution." J. Nonparametr. Statist., vol. 7, pp. 307–330.

Passow, E. (1977). "Polynomials with positive coefficients: uniqueness of best approximation." J. Approximation Theory, vol. 21, pp. 352–355.

Redner, R. A. and Walker, H. F. (1984). "Mixture densities, maximum likelihood and the EM algorithm." SIAM Rev., vol. 26, pp. 195–239.

Stefanski, L. and Carroll, R. J. (1990). "Deconvoluting kernel density estimators." Statistics, vol. 21, pp. 169–184.

Wang, X.-F. and Wang, B. (2011). "Deconvolution estimation in measurement error models: The R package decon." Journal of Statistical Software, vol. 39, pp. 1–24.

Wang, X.-F. and Ye, D. (2015). "Conditional density estimation in measurement error problems." Journal of Multivariate Analysis, vol. 133, pp. 38 – 50.

Wu, C.-F. J. (1983). "On the convergence properties of the EM algorithm." Ann. Statist., vol. 11, pp. 95–103.

Zhang, C.-H. (1990). "Fourier methods for estimating mixing densities and distributions." Ann. Statist., vol. 18, pp. 806–831.